TRƯỜNG ĐẠI HỌC FPT

CƠ SỞ TP HỒ CHÍ MINH

Phân tích và kiểm định thống kê dữ liệu thời tiết bằng Python

Môn học: MAS291 - Statistics & Probability

Sinh viên thực hiện:

Trần Kiến Quốc SE190250

Vũ Đức SE192458

Võ Kế Trí SE196402

Mai Minh Quân SE204020

Phạm Tùng Dương SE203523

Trần Minh Quý SE196687

Giảng viên hướng dẫn:

Nguyễn Trần Minh Thư

Ngày 3 tháng 11 năm 2025

Mục lục

1 Giới thiệu

Thời tiết đóng vai trò quan trọng trong đời sống hằng ngày và ảnh hưởng đến nhiều lĩnh vực như nông nghiệp, sức khỏe, giao thông và tiêu thụ năng lượng. Việc hiểu rõ các mẫu hình thời tiết và mối quan hệ giữa các yếu tố khí tượng là cần thiết cho cả nghiên cứu khoa học và ứng dụng thực tiễn.

Trong dự án này, nhóm tiến hành phân tích một bộ dữ liệu thứ cấp về thời tiết, bao gồm các biến số như nhiệt độ, nhiệt độ cảm nhận, độ ẩm, tốc độ gió, hướng gió, tầm nhìn, độ che phủ mây và áp suất khí quyển. Dữ liệu được xử lý và khám phá bằng ngôn ngữ Python, thông qua các kỹ thuật thống kê mô tả và trực quan hóa dữ liệu nhằm tóm tắt những đặc điểm chính của tập dữ liệu.

Các phương pháp thống kê suy luận như kiểm định giả thuyết và ước lượng khoảng tin cậy được áp dụng để rút ra những kết luận có ý nghĩa về tập dữ liệu tổng thể. Mục tiêu của nghiên cứu là minh họa cách tư duy thống kê có thể được sử dụng để chuyển đổi dữ liệu thô về thời tiết thành các thông tin hữu ích phục vụ cho phân tích và ra quyết định.

2 Thu thập và xử lý dữ liệu

2.1 Thu thập dữ liệu

Tập dữ liệu thời tiết được tải trực tiếp từ nguồn dữ liệu thứ cấp có sẵn trên Kaggle.

Tập dữ liệu bao gồm các cột thông tin:

- **Formatted Date**: Ngày và giờ quan sát được định dạng sẵn.
- **Summary**: Mô tả ngắn gọn về điều kiện thời tiết trong ngày.
- **Precip Type**: Loại giáng thủy (mưa hoặc tuyết).
- **Temperature (C)**: Nhiệt độ trung bình trong ngày (đơn vị: độ C).
- **Apparent Temperature (C)**: Nhiệt độ cảm nhận được (độ C).
- **Humidity**: Độ ẩm tương đối của không khí (từ 0 đến 1).
- **Wind Speed (km/h)**: Tốc độ gió trung bình (km/h).

- **Wind Bearing (degrees)**: Hướng gió theo độ.
- **Visibility (km)**: Tầm nhìn xa trung bình trong ngày.
- **Cloud Cover**: Mức độ che phủ mây (từ 0 đến 1).
- **Pressure (millibars)**: Áp suất khí quyển trung bình (millibar).
- **Daily Summary**: Mô tả tổng quát về điều kiện thời tiết trong ngày.

2.2 Xử lý dữ liệu

2.2.1 Xử lý giá trị thiếu và loại bỏ trùng lặp

```
print(df.isna().sum())


df = df.dropna()
df = df.drop_duplicates()
```

Listing 1: Xử lý giá trị thiếu và loại bỏ trùng lặp

2.2.2 Chuyển đổi kiểu dữ liệu

```
df['Formatted Date'] = pd.to_datetime(df['Formatted Date'], utc=True)
```

Listing 2: Chuyển đổi kiểu dữ liệu

3 Thống kê mô tả

3.1 Khởi tạo và phân loại biến

Mục tiêu: xác định rõ những cột thuộc dạng **qualitative** (định tính) và **quantitative** (định lượng) để áp dụng các kỹ thuật phù hợp.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
import seaborn as sns

df = pd.read_csv('dataset/weatherHistory_cleaned.csv')

qualitative_cols = ['Summary', 'Precip Type']
quantitative_cols = [
    'Temperature (C)', 'Apparent Temperature (C)', 'Humidity',
    'Wind Speed (km/h)', 'Wind Bearing (degrees)',
    'Visibility (km)', 'Pressure (millibars)'
]
```

Listing 3: Khởi tạo và phân loại cột

Việc phân loại chính xác giúp lựa chọn biểu đồ (pie/bar cho định tính; histogram/boxplot cho định lượng) và lựa chọn các thống kê phù hợp.

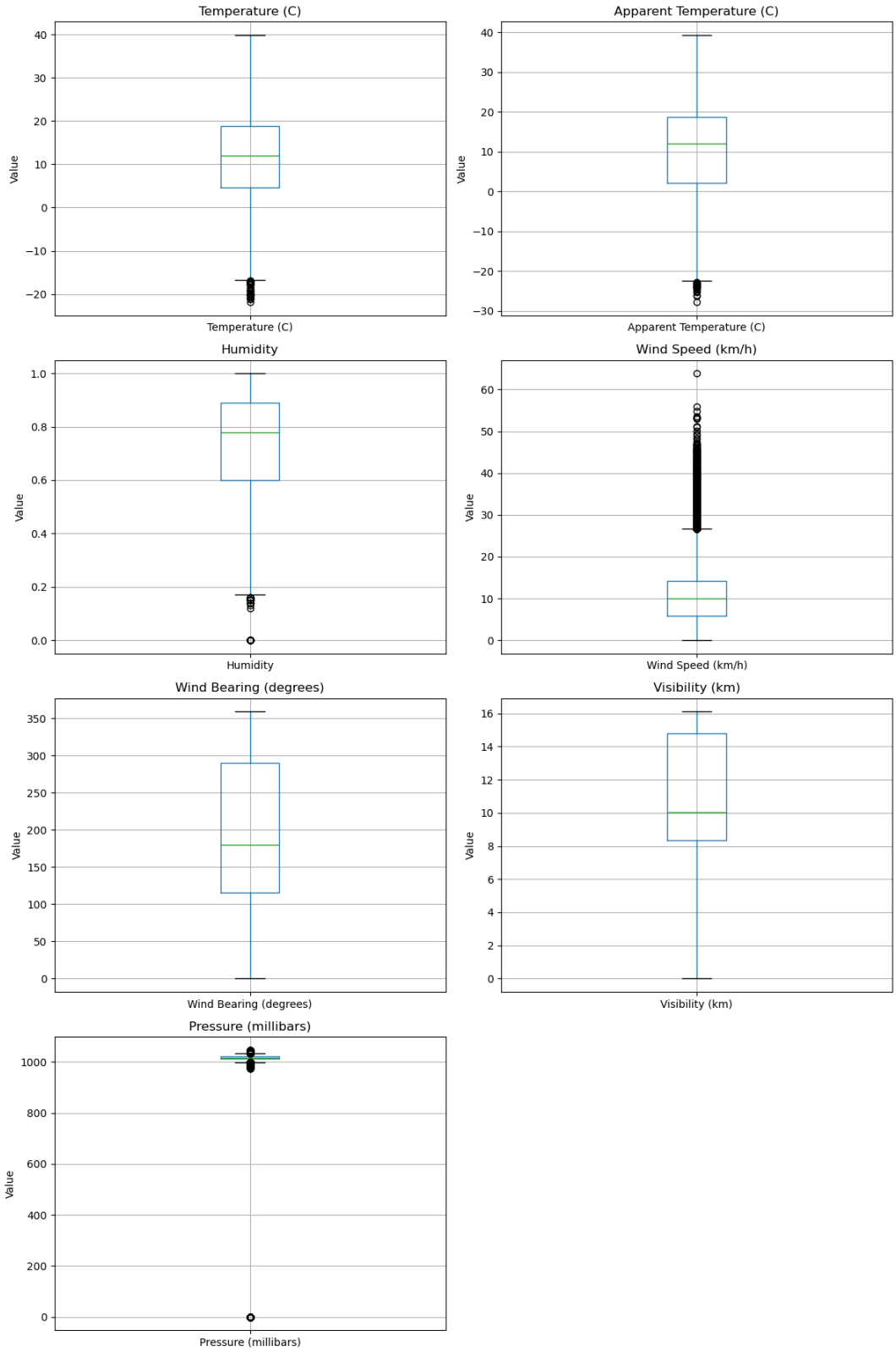
3.2 Phát hiện ngoại lệ

Nguyên lý phương pháp IQR:

- Tính **Q1** (tứ phân vị thứ nhất – 25%) và **Q3** (tứ phân vị thứ ba – 75%).
- Tính $\mathbf{IQR = Q3 - Q1}$.
- Một điểm dữ liệu được coi là ngoại lai nếu:

$$x < Q1 - 1.5 \times IQR \quad \text{hoặc} \quad x > Q3 + 1.5 \times IQR$$

Hình 1 trình bày các biểu đồ hộp (boxplot) cho các biến định lượng trong tập dữ liệu. Biểu đồ hộp giúp trực quan hoá phân bố dữ liệu, giá trị trung vị (median), tứ phân vị (Q1, Q3) và đặc biệt là các điểm ngoại lệ (outliers).



Hình 1: Boxplot của các biến định lượng trong tập dữ liệu thời tiết

Nhìn chung, hầu hết các biến số khí hậu đều thể hiện sự ổn định tương đối, song vẫn tồn tại một

số giá trị ngoại lệ tại các vùng phân phối. Các biến **Nhiệt Độ** (Temperature) và **Nhiệt Độ Cảm Nhận** (Apparent Temperature) có một lượng nhỏ ngoại lệ ở vùng giá trị **thấp** (dưới $-10°C$), phản ánh các thời điểm lạnh bất thường nhưng vẫn nằm trong phạm vi dữ liệu thực tế. Ngược lại, biến **Độ Ẩm** (Humidity) và **Hướng Gió** (Wind Bearing) gần như **không có** ngoại lệ rõ ràng; phân phối của Độ Ẩm ổn định quanh mức cao (0.7–1.0), trong khi Hướng Gió phân bố đồng đều trong khoảng 0° đến 360°. Giá trị ngoại lệ rõ rệt nhất nằm ở **Tốc Độ Gió** (Wind Speed), với một vài điểm ở **phía trên** (trên 25 km/h) đại diện cho các sự kiện gió mạnh hoặc bão. Tương tự, **Tầm Nhìn** (Visibility) cũng có ngoại lệ ở vùng **thấp** (dưới 2 km) do sương mù dày hoặc mưa lớn, dù phần lớn dữ liệu tập trung ở 8–16 km. Cuối cùng, **Áp Suất Khí Quyển** (Pressure) có một số ngoại lệ ở **cả hai đầu** (thấp $\approx$ 1000 mb và cao $>$ 1030 mb), song phân phối chung vẫn khá ổn định quanh giá trị trung vị 1016 mb.

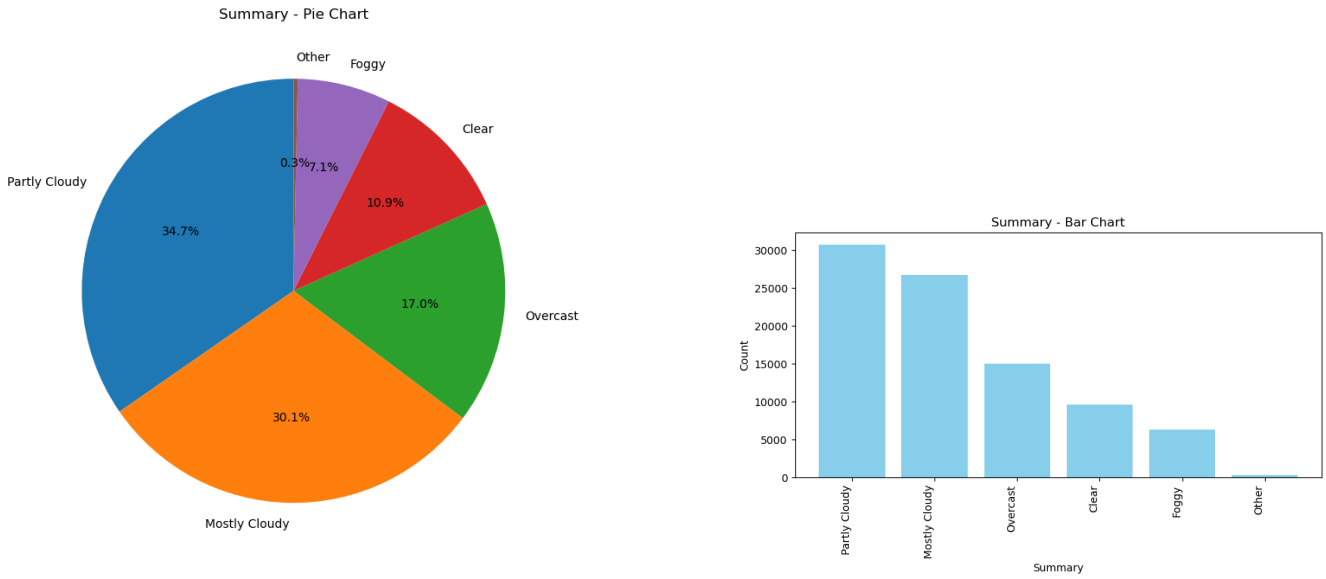
3.3 Dữ liệu định tính (Qualitative Data)

3.3.1 Phân tích biến Summary

Biến **Summary** thể hiện tình trạng thời tiết tổng quát trong tập dữ liệu. Bảng 1 dưới đây trình bày tần suất xuất hiện và tỷ lệ phần trăm của các trạng thái thời tiết phổ biến nhất.

Bảng 1: Tần suất và tỷ lệ phần trăm của các trạng thái thời tiết (Summary)

Trạng thái thời tiết	**Số lần xuất hiện (Count)**	**Tỷ lệ (%)**
Partly Cloudy	30,739	34.65
Mostly Cloudy	26,704	30.10
Overcast	15,044	16.96
Clear	9,635	10.86
Foggy	6,312	7.11
Other	287	0.32



(a) Biểu đồ tròn (Pie chart)

(b) Biểu đồ cột (Bar chart)

Hình 2: Phân bố các trạng thái thời tiết trong biến **Summary**

Nhận xét

Từ bảng và các biểu đồ trên, có thể thấy rằng các trạng thái **Partly Cloudy** (mây rải rác) và **Mostly Cloudy** (nhiều mây) chiếm tỷ trọng lớn nhất, lần lượt là **34.7%** và **30.1%**. Điều này cho thấy phần lớn thời gian trong giai đoạn quan sát, bầu trời thường xuất hiện mây ở các mức độ khác nhau.

Các trạng thái **Overcast** (u ám) và **Clear** (trời quang) có tỉ lệ lần lượt là **17.0%** và **10.9%**, phản ánh sự xen kẽ giữa những ngày âm u và những ngày nắng rõ. Trong khi đó, hiện tượng **Foggy** (sương mù) xuất hiện ít hơn, chỉ chiếm khoảng **7.1%**, và nhóm **Other** chỉ chiếm **0.3%**.

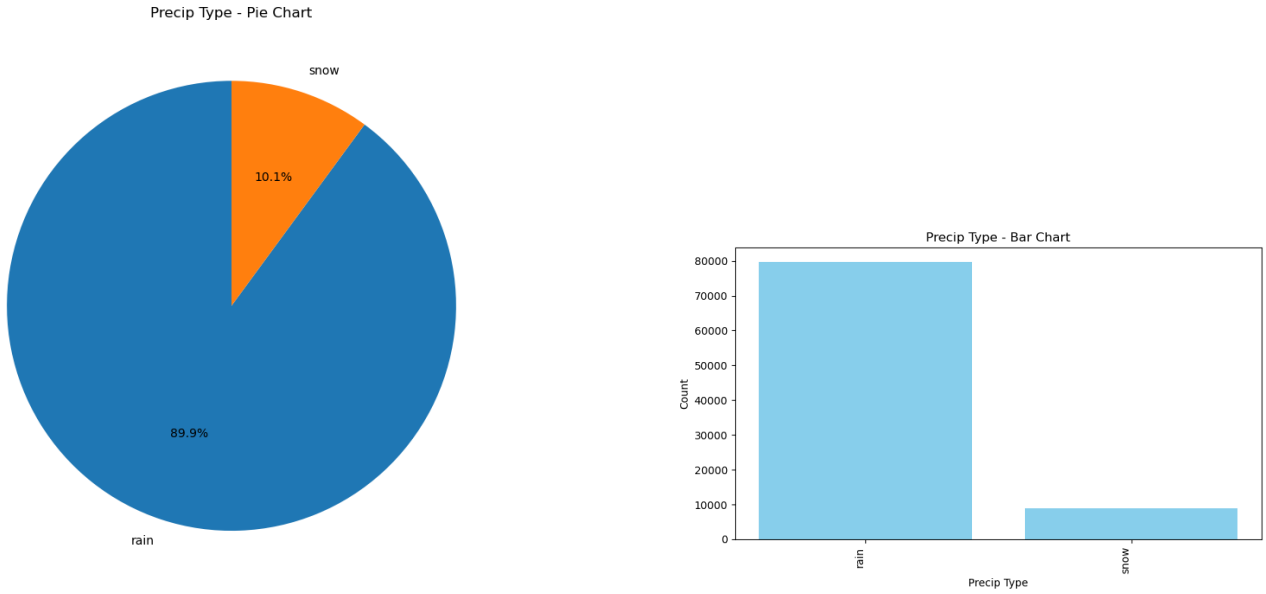
Tổng thể, dữ liệu cho thấy thời tiết tại khu vực quan sát có xu hướng **ôn hòa, ít biến động mạnh và thường xuyên có mây**, phù hợp với đặc điểm của khí hậu ổn định và ôn đới.

3.3.2 Phân tích biến Precip Type

Biến **Precip Type** mô tả loại hình giáng thủy (mưa, tuyết) ghi nhận trong dữ liệu thời tiết. Bảng 2 trình bày tần suất và tỷ lệ phần trăm của từng loại giáng thủy.

Bảng 2: Tần suất và tỷ lệ phần trăm của các loại giáng thủy (Precip Type)

Loại giáng thủy	**Số lần xuất hiện (Count)**	**Tỷ lệ (%)**
Rain	79,772	89.91
Snow	8,949	10.09



(a) Biểu đồ tròn (Pie chart)

(b) Biểu đồ cột (Bar chart)

Hình 3: Phân bố các loại giáng thủy trong biến **Precip Type**

Nhận xét

Quan sát từ bảng và các biểu đồ cho thấy loại giáng thủy chiếm ưu thế là **Rain** (mưa) với tỷ lệ **89.91%**, trong khi **Snow** (tuyết) chỉ chiếm **10.09%**. Điều này phản ánh rõ ràng đặc trưng khí hậu của khu vực nghiên cứu — nơi thời tiết chủ yếu là mưa và chỉ có tuyết xuất hiện trong một số giai đoạn ngắn, thường là vào mùa đông.

Tỷ lệ mưa áp đảo cho thấy đây là **khu vực có khí hậu ẩm, ôn hòa và ít chịu ảnh hưởng của lạnh cực độ**, phù hợp với các vùng ven biển hoặc có khí hậu cận nhiệt đới. Ngược lại, tuyết xuất hiện với tần suất thấp cho thấy điều kiện thời tiết khắc nghiệt như băng giá chỉ xảy ra trong những thời điểm giới hạn.

3.4 Dữ liệu định lượng (Quantitative Data)

3.4.1 Thống kê mô tả dữ liệu định lượng

Nhằm hiểu rõ hơn đặc trưng tổng thể của các biến số thực trong tập dữ liệu, nhóm tiến hành thống kê mô tả (Descriptive Statistics) cho toàn bộ các biến định lượng: *Temperature (C)*, *Apparent Temperature (C)*, *Humidity*, *Wind Speed (km/h)*, *Wind Bearing (degrees)*, *Visibility (km)*, và *Pressure (millibars)*.

Bảng 3: Thống kê mô tả cho các biến định lượng

Biến	Count	Mean	Std	Min	25%	50%	75%	Max
Temperature (C)	88,721	12.37	9.47	-16.71	5.00	12.32	19.03	38.98
Apparent Temperature (C)	88,721	11.39	10.56	-22.42	2.93	12.32	19.03	39.34
Humidity	88,721	0.73	0.20	0.17	0.60	0.78	0.90	1.00
Wind Speed (km/h)	88,721	10.14	5.84	0.00	5.57	9.66	13.83	26.61
Wind Bearing (degrees)	88,721	186.55	107.27	0.00	113.00	180.00	289.00	359.00
Visibility (km)	88,721	10.47	4.15	0.00	8.89	10.05	14.91	16.10
Pressure (millibars)	88,721	1,016.76	6.74	998.17	1,012.33	1,016.55	1,020.90	1,034.79

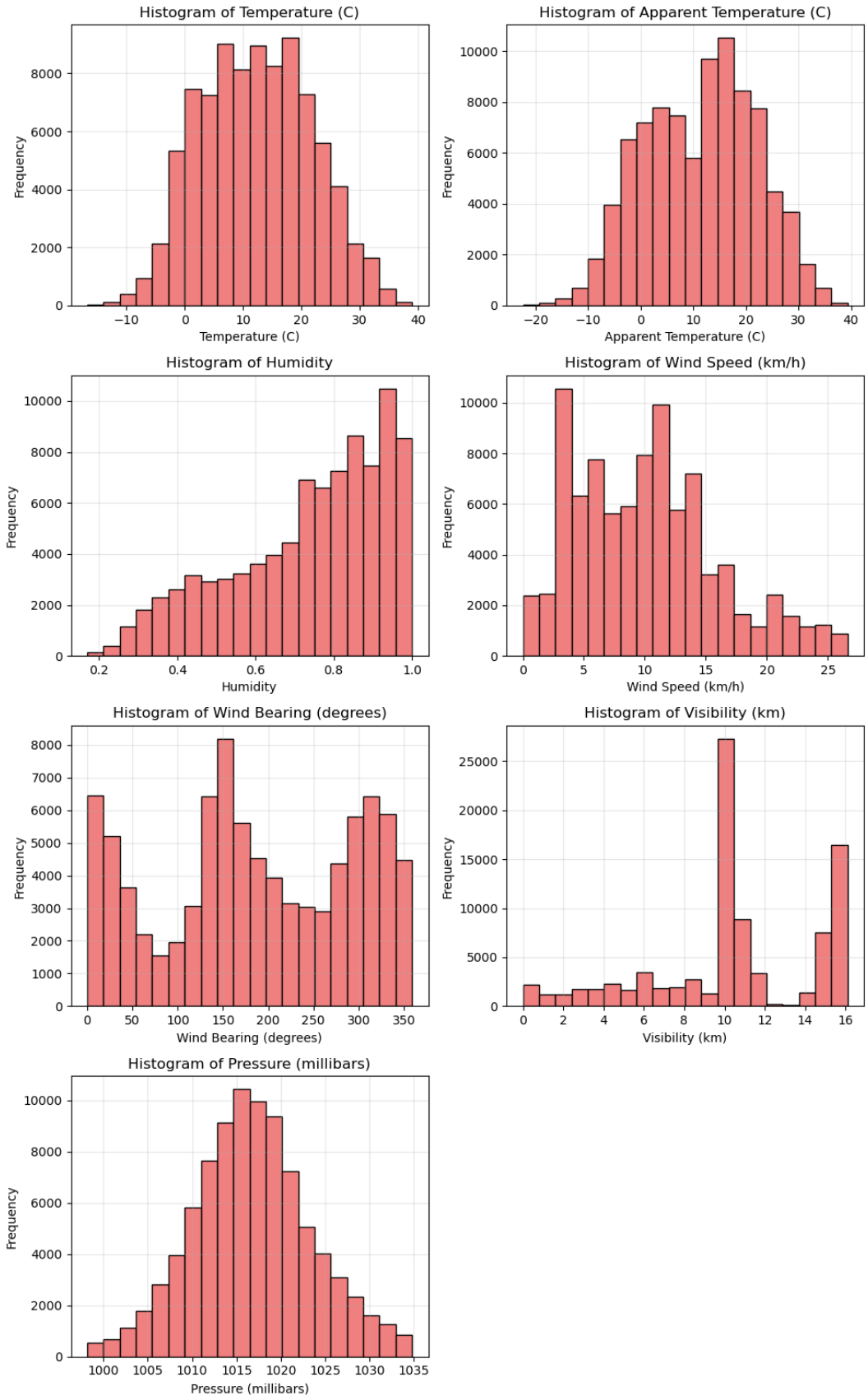
Nhận xét

- Nhiệt độ trung bình đạt khoảng **12.37°C**, dao động từ **-16.71°C** đến **38.98°C**, cho thấy dữ liệu bao quát đủ cả bốn mùa trong năm.
- Nhiệt độ cảm nhận (*Apparent Temperature*) có trung bình thấp hơn một chút (**11.39°C**), phản ánh ảnh hưởng của gió và độ ẩm đến cảm giác nhiệt thực tế.
- Độ ẩm trung bình **0.73** (tương đương 73%) cho thấy khí hậu nhìn chung khá ẩm, đặc trưng của vùng có lượng mưa cao.
- Tốc độ gió trung bình đạt **10.14 km/h**, chủ yếu theo hướng **186°** (hướng Nam), cho thấy sự ổn định của hướng gió trong khu vực.
- Tầm nhìn trung bình **10.47 km** — khá tốt — nhưng giá trị nhỏ nhất bằng 0 cho thấy có thời điểm xuất hiện sương mù dày đặc hoặc mưa lớn.
- Áp suất khí quyển trung bình **1,016.76 millibars**, dao động trong phạm vi bình thường (**998.17–1,034.79**), phản ánh điều kiện khí quyển ổn định.

Tổng thể, các biến định lượng cho thấy **khí hậu ôn hòa, độ ẩm cao, và có sự biến thiên theo mùa rõ rệt**. Dữ liệu không xuất hiện giá trị bất thường đáng kể và thể hiện tốt đặc trưng của một vùng khí hậu ổn định, thường xuyên có mây và mưa.

3.4.2 Phân phối của các biến định lượng

Hình 4 mô tả phân phối của các biến định lượng trong tập dữ liệu thời tiết. Các biểu đồ histogram giúp quan sát xu hướng và đặc điểm của từng biến, từ đó hỗ trợ lựa chọn phương pháp phân tích phù hợp.



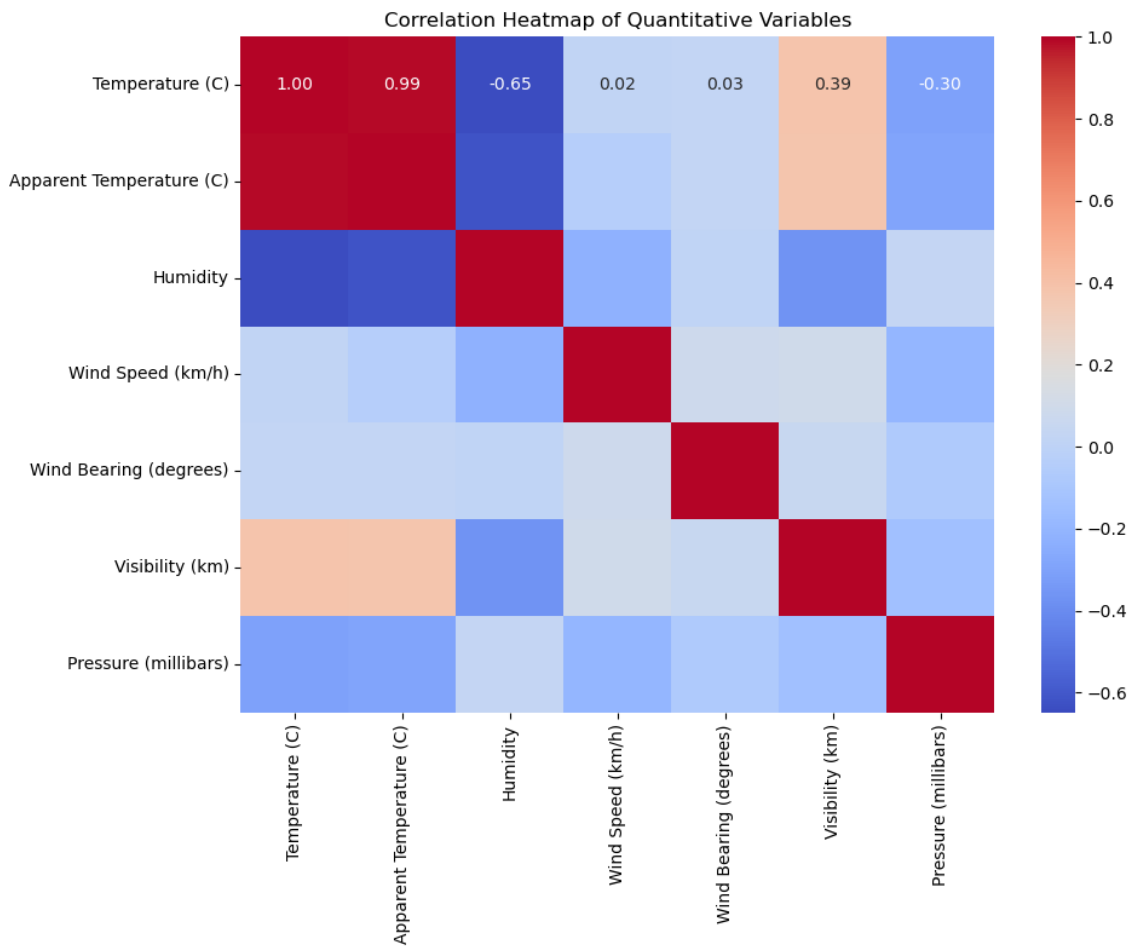
Hình 4: Phân phối của các biến định lượng trong tập dữ liệu thời tiết

Các biến **Nhiệt Độ** (Temperature) và **Nhiệt Độ Cảm Nhận** (Apparent Temperature) có phân phối **xấp xỉ chuẩn** ($\approx \mathcal{N}$), tập trung quanh mức $\mu \approx 12°\text{C}$ với dao động nhỏ (chủ yếu $0°\text{C} - 25°\text{C}$). Apparent Temperature thường **thấp hơn** T do tác động của gió và độ ẩm. Ngược lại, **Độ Ẩm** (Humidity) thể hiện phân phối **nghiêng phải**, với phần lớn quan sát nằm trong khoảng **cao** (0.6–1.0), phù hợp với đặc trưng khí hậu nhiệt đới.

Đối với **Tốc Độ Gió** (Wind Speed), phân phối cũng **hơi lệch phải**, cho thấy tốc độ gió **trung bình** (5–10 km/h) chiếm ưu thế, song vẫn ghi nhận các giá trị cao (> 25 km/h) biểu thị gió mạnh. **Hướng Gió** (Wind Bearing) có phân phối **tương đối đồng đều** nhưng nổi bật ở một số hướng **chính** ($0°, 180°, 270°$).

Về **Tầm Nhìn** (Visibility), phân phối **lệch phải rõ rệt**, cho thấy điều kiện **tầm nhìn xa** (10–16 km) là phổ biến, mặc dù tồn tại những thời điểm có tầm nhìn thấp do sương mù hoặc mưa lớn. Cuối cùng, **Áp Suất Khí Quyển** (Pressure) có phân phối **gần chuẩn** ($\mu \approx 1016$ mb), dao động nhẹ, thể hiện sự ổn định tương đối trong điều kiện khí áp.

3.5 Phân Tích Mối Tương Quan Giữa Các Biến Định Lượng



Hình 5: Biểu đồ tương quan giữa các biến định lượng

Phân tích biểu đồ nhiệt cho thấy mối tương quan mạnh mẽ nhất là giữa **Temperature** và **Apparent Temperature** ($r = 0.99$), cho thấy tính đồng nhất cao, cần lưu ý về vấn đề đa cộng tuyến. Cặp tương quan mạnh và đáng chú ý tiếp theo là mối quan hệ **nghịch chiều mạnh** giữa Nhiệt độ (cả T và $T_{Apparent}$) và **Humidity** ($r \approx -0.65$). Ngoài ra, Nhiệt độ có tương quan **thuận chiều trung bình** với **Visibility** ($r \approx 0.38$). Ngược lại, các biến **Wind Speed** và **Wind Bearing** gần như **độc lập** với các biến còn lại (tương quan rất yếu, $|r| \leq 0.02$). Tóm lại, các yếu tố nhiệt độ và độ ẩm là các biến có mối quan hệ phụ thuộc rõ rệt nhất trong bộ dữ liệu, là trọng tâm khi xây dựng mô hình dự báo.

Tính giá trị tới hạn (Critical Value)

Để xác định giá trị tới hạn khi xây dựng khoảng tin cậy (Confidence Interval) ở mức 95%, ta thực hiện các bước sau:

- Mức tin cậy:

$$\text{confidence\_level} = 0.95$$

- Mức ý nghĩa:

$$\alpha = 1 - \text{confidence\_level} = 0.05$$

- Xác định giá trị biên của phân phối chuẩn hoặc phân phối t:

$$\text{ppf\_input} = 1 - \frac{\alpha}{2} = 0.975$$

- Khi đó, giá trị tới hạn được tính bằng:

$$z_{\text{crit}} = z_{0.975} = \texttt{norm.ppf(ppf\_input)}$$

$$t_{\text{crit}} = t_{0.975,\, df=n-1} = \texttt{t.ppf(ppf\_input, df=n-1)}$$

Các giá trị này tương ứng với biên của khoảng tin cậy 95% cho phân phối chuẩn và phân phối t.

Triển khai

```
confidence_level = 0.95         # 95% CI
alpha = 1 - confidence_level    # 0.05
ppf_input = 1 - alpha / 2       # 0.975

t_crit = t.ppf(ppf_input, df=n-1)
z_crit = norm.ppf(ppf_input)
```

4 Kiểm định giả thuyết một biến mẫu

4.1 Lý thuyết và công thức toán học

4.1.1 Giới thiệu về kiểm định giả thuyết

Kiểm định giả thuyết (Hypothesis Testing) là một phương pháp thống kê dùng để đưa ra kết luận về một đặc trưng (parameter) của quần thể dựa trên dữ liệu mẫu. Quá trình này gồm hai giả thuyết đối lập:

- **Giả thuyết không (Null hypothesis, H0)**: là giả thuyết cơ bản, thường biểu thị rằng không có sự khác biệt hoặc không có hiệu ứng.
- **Giả thuyết đối (Alternative hypothesis, H1)**: là giả thuyết mà chúng ta muốn chứng minh.

Các bước chung của kiểm định:

1. Chọn thống kê kiểm định phù hợp với loại dữ liệu và giả thuyết.
2. Tính giá trị thống kê (t-statistic, z-statistic, ...).
3. Xác định p-value: xác suất quan sát giá trị thống kê như đã tính, nếu H0 đúng.
4. So sánh p-value với mức ý nghĩa α (thường 0.05) để quyết định bác bỏ hay không bác bỏ H0.

4.1.2 Kiểm định trung bình một biến

Giả sử $X_1, X_2, ..., X_n$ là mẫu ngẫu nhiên từ quần thể với trung bình μ và phương sai σ^2. Chúng ta muốn kiểm tra H0: $\mu = \mu_0$ với một giá trị giả định μ_0.

- Trung bình mẫu:
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
- Độ lệch chuẩn mẫu:
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- Thống kê t:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

- Khoảng tin cậy cho trung bình quần thể:

$$\text{CI} = \bar{X} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

4.1.3 Kiểm định tỷ lệ một biến

Giả sử một biến nhị phân (binary) với xác suất thành công p. Chúng ta có mẫu gồm n quan sát và số "thành công" x.

- Tỷ lệ mẫu:

$$\hat{p} = \frac{x}{n}$$

- Giả thuyết H0: $p = p_0$ với giá trị giả định p_0.

- Thống kê z:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

  Phân phối chuẩn chuẩn hóa: $z \sim N(0,1)$.

- Khoảng tin cậy cho tỷ lệ quần thể (mức tin cậy $1-\alpha$):

$$\text{CI} = \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

4.1.4 Diễn giải

- Nếu p-value $< \alpha$, bác bỏ H0, nghĩa là dữ liệu cung cấp bằng chứng thống kê để chấp nhận H1.
- Nếu p-value $\geq \alpha$, không bác bỏ H0, nghĩa là không có bằng chứng đủ để kết luận H1.
- Khoảng tin cậy cung cấp phạm vi các giá trị hợp lý cho tham số quần thể dựa trên dữ liệu mẫu.

4.2 Kiểm định trung bình nhiệt độ mùa hè

4.2.1 Giả thuyết

Giả thuyết được đặt ra nhằm kiểm tra xem **nhiệt độ trung bình mùa hè có khác 20°C hay không**.

$$H_0 : \mu = 20$$

$$H_1 : \mu \neq 20$$

4.2.2 Code Python

```
summer_df = df[df['Formatted Date'].dt.month.isin([6, 7, 8])]
temps_summer = summer_df['Temperature (C)']

t_stat, p_two_tailed = stats.ttest_1samp(temps_summer, 20)

mean_temp = temps_summer.mean()
std_temp = temps_summer.std(ddof=1)

n = len(temps_summer)
se = std_temp / np.sqrt(n)

ci_lower = mean_temp - t_crit * se
ci_upper = mean_temp + t_crit * se
```

4.2.3 Kết quả

- t-statistic = 52.845, two-tailed p-value = 0.000
- 95% confidence interval = (21.900, 22.085)

4.2.4 Nhận xét

- Trung bình nhiệt độ mùa hè nằm trong khoảng 21.90–22.09°C.

- Với p-value < 0.05, ta **bác bỏ giả thuyết H0**.
- Kết luận: Nhiệt độ trung bình mùa hè thực tế cao hơn 20°C với mức ý nghĩa thống kê 95%.

4.3 Kiểm định tỷ lệ ngày có tuyết mùa đông

4.3.1 Giả thuyết

Giả thuyết được đặt ra nhằm kiểm tra xem **tỷ lệ ngày có tuyết trong mùa đông có vượt quá 37.4% hay không**.

$$H_0 : p \leq 0.374$$

$$H_1 : p > 0.374$$

4.3.2 Code Python

```
value = 0.374

winter_df = df[df['Formatted Date'].dt.month.isin([12, 1, 2])]
n_total = len(winter_df)
n_snow = (winter_df['Precip Type'] == 'snow').sum()

count = n_snow
nobs = n_total

stat, p_two_tailed = proportions_ztest(count, nobs, value)
p_one_tailed = p_two_tailed / 2 if stat > 0 else 1 - (p_two_tailed / 2)

p_hat = n_snow / n_total
se_diff = np.sqrt(p_hat * (1 - p_hat) / n_total)

ci_lower = p_hat - z_crit * se_diff
ci_upper = p_hat + z_crit * se_diff
```

4.3.3 Kết quả

- z-statistic = -1.872, one-tailed p-value = 0.969
- 95% confidence interval = (0.361, 0.374)

4.3.4 Nhận xét

- Tỷ lệ ngày có tuyết mùa đông ước tính khoảng 36.1–37.4%, với tỷ lệ mẫu là 36.8%.
- Kiểm định one-tailed với $H_0 : p \leq 0.374$ cho thấy p-value = $0.969 > 0.05$, **không bác bỏ H0**.
- Như vậy, theo dữ liệu, không có đủ bằng chứng thống kê để kết luận rằng tỷ lệ ngày có tuyết mùa đông vượt quá 0.374.

5 Kiểm định giả thuyết hai biến mẫu

5.1 Lý thuyết và công thức toán học

5.1.1 Giới thiệu

Kiểm định 2 mẫu dùng để so sánh đặc trưng (mean hoặc proportion) giữa hai quần thể. Giả sử ta có hai mẫu độc lập:

$$X_1, X_2, ..., X_{n_1} \sim \text{Population 1}, \quad Y_1, Y_2, ..., Y_{n_2} \sim \text{Population 2}$$

Các bước kiểm định tương tự 1 mẫu:

1. Xác định H0 và H1.
2. Chọn thống kê kiểm định phù hợp.
3. Tính giá trị thống kê và p-value.
4. So sánh p-value với mức ý nghĩa α.

5.1.2 Kiểm định trung bình hai quần thể

- Trung bình mẫu:

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \quad \bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

- Độ lệch chuẩn mẫu:

$$s_X = \sqrt{\frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2}, \quad s_Y = \sqrt{\frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2}$$

- Pooled standard deviation (giả sử phương sai bằng nhau):

$$s_p = \sqrt{\frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}}$$

- Thống kê t:

$$t = \frac{\bar{X} - \bar{Y}}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad t \sim t_{n_1+n_2-2}$$

- Khoảng tin cậy cho hiệu trung bình:

$$\text{CI} = (\bar{X} - \bar{Y}) \pm t_{\alpha/2, n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

5.1.3 Kiểm định tỷ lệ hai quần thể

- Tỷ lệ mẫu:

$$\hat{p}_1 = \frac{x_1}{n_1}, \quad \hat{p}_2 = \frac{x_2}{n_2}$$

- Giả thuyết H0: $p_1 = p_2$, H1: $p_1 \neq p_2$

- Thống kê z:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}}, \quad z \sim N(0, 1)$$

- Khoảng tin cậy 95% cho hiệu tỷ lệ:

$$\text{CI} = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

5.1.4 Diễn giải

- Nếu p-value $< \alpha$, bác bỏ H0: có sự khác biệt thống kê giữa hai quần thể.
- Nếu p-value $\geq \alpha$, không bác bỏ H0: không đủ bằng chứng cho sự khác biệt.
- Khoảng tin cậy cung cấp phạm vi giá trị hợp lý cho hiệu trung bình hoặc hiệu tỷ lệ.

5.2 Kiểm định nhiệt độ trung bình

Giả thuyết được đặt ra nhằm kiểm tra xem **nhiệt độ trung bình trong ngày "Mostly Cloudy" có cao hơn so với "Partly Cloudy" hay không**.

5.2.1 Giả thuyết

$$H_0 : \mu_{\text{Mostly Cloudy}} - \mu_{\text{Partly Cloudy}} \leq 0$$
$$H_1 : \mu_{\text{Mostly Cloudy}} - \mu_{\text{Partly Cloudy}} > 0$$

5.2.2 Code Python

```
mostly_cloudy = df[df['Summary'] == 'Mostly Cloudy']['Temperature (C)']
partly_cloudy = df[df['Summary'] == 'Partly Cloudy']['Temperature (C)']

t_stat, p_two_tailed = stats.ttest_ind(mostly_cloudy, partly_cloudy,
    ↪ equal_var=True)

p_one_tailed = p_two_tailed / 2 if t_stat > 0 else 1 - (p_two_tailed / 2)

n1, n2 = len(mostly_cloudy), len(partly_cloudy)
mean1, mean2 = mostly_cloudy.mean(), partly_cloudy.mean()
```

```
std1, std2 = mostly_cloudy.std(ddof=1), partly_cloudy.std(ddof=1)
s_pooled = np.sqrt(((n1-1)*std1**2 + (n2-1)*std2**2) / (n1+n2-2))
se_diff = s_pooled * np.sqrt(1/n1 + 1/n2)

dfree = n1 + n2 - 2
ci_lower = (mean1 - mean2) - t_crit * se_diff
ci_upper = (mean1 - mean2) + t_crit * se_diff
```

5.2.3 Kết quả

- t-statistic = -45.223
- one-tailed p-value = 1.000
- 95% CI cho hiệu trung bình (Mostly - Partly) = (-3.562, -3.197)
- Trung bình nhiệt độ: Mostly Cloudy = 12.85°C, Partly Cloudy = 16.23°C

5.2.4 Nhận xét

- Trung bình nhiệt độ nhóm **Mostly Cloudy** thấp hơn nhóm **Partly Cloudy** khoảng 3.20–3.56°C.
- Giá trị p-value = 1.000 $>$ 0.05 cho thấy không có đủ bằng chứng để bác bỏ giả thuyết H_0.
- Kết luận: Nhiệt độ của nhóm **Mostly Cloudy** không cao hơn nhóm **Partly Cloudy**, trái với giả thuyết ban đầu đề ra.

5.3 Kiểm định tỷ lệ mưa

5.3.1 Giả thuyết

Giả thuyết được đặt ra nhằm kiểm tra xem **tỷ lệ mưa trong các ngày "Mostly Cloudy" có khác so với "Partly Cloudy" hay không**.

$$H_0 : p_{\text{Mostly Cloudy}} = p_{\text{Partly Cloudy}}$$

$$H_a : p_{\text{Mostly Cloudy}} \neq p_{\text{Partly Cloudy}}$$

5.3.2 Code Python

```
mostly_cloudy = df[df['Summary'] == 'Mostly Cloudy']
partly_cloudy = df[df['Summary'] == 'Partly Cloudy']

count1 = (mostly_cloudy['Precip Type'] == 'rain').sum()
nobs1 = len(mostly_cloudy)

count2 = (partly_cloudy['Precip Type'] == 'rain').sum()
nobs2 = len(partly_cloudy)

count = np.array([count1, count2])
nobs = np.array([nobs1, nobs2])

stat, p_two_tailed = proportions_ztest(count, nobs)
p_value = p_two_tailed

p1_hat = count1 / nobs1
p2_hat = count2 / nobs2
diff = p1_hat - p2_hat

se_diff = np.sqrt(p1_hat*(1-p1_hat)/nobs1 + p2_hat*(1-p2_hat)/nobs2)

ci_lower = diff - z_crit * se_diff
ci_upper = diff + z_crit * se_diff
```

5.3.3 Kết quả

- Tỷ lệ mưa: Mostly Cloudy = 0.937, Partly Cloudy = 0.951
- z-statistic = -7.125
- two-tailed p-value = 0.000
- 95% CI cho hiệu tỷ lệ (Mostly - Partly) = (-0.017, -0.010)

5.3.4 Nhận xét và kết luận

- Nhóm **Mostly Cloudy** có tỷ lệ mưa thấp hơn nhóm **Partly Cloudy** khoảng 1.0–1.7%.
- p-value = $0.000 < 0.05$, bác bỏ giả thuyết H_0.
- Kết luận: Có sự khác biệt có ý nghĩa thống kê về tỷ lệ mưa giữa hai nhóm **Mostly Cloudy** và **Partly Cloudy**.

6 Kết Luận

Mặc dù tập dữ liệu sử dụng trong nghiên cứu là dữ liệu thứ cấp, được thu thập trong giai đoạn 2006–2016, các kết quả phân tích vẫn mang giá trị minh họa rõ ràng cho quy trình xử lý và khai thác dữ liệu bằng các phương pháp thống kê hiện đại. Báo cáo đã cho thấy cách kết hợp hiệu quả giữa ngôn ngữ lập trình Python và các kỹ thuật thống kê trong việc hiểu, trực quan hóa và kiểm định các giả thuyết liên quan đến hiện tượng tự nhiên.

Thông qua quá trình phân tích dữ liệu thời tiết, nhóm đã áp dụng thành công các phương pháp thống kê mô tả và kiểm định giả thuyết để làm rõ đặc trưng khí hậu của tập dữ liệu. Kết quả chỉ ra rằng thời tiết nhìn chung ổn định, chủ yếu có mây và mưa, phù hợp với đặc điểm của vùng khí hậu ôn hòa. Các kiểm định thống kê giúp làm sáng tỏ những khác biệt có ý nghĩa về nhiệt độ giữa các điều kiện thời tiết, cũng như sự phân bố của mưa và tuyết theo mùa.

Bên cạnh giá trị học thuật, nghiên cứu này còn minh họa quy trình phân tích dữ liệu khoa học từ tiền xử lý, mô tả, trực quan đến kiểm định giả thuyết. Quy trình này có thể được mở rộng và áp dụng cho các bộ dữ liệu thời tiết mới hơn hoặc dữ liệu thực tế từ các trạm quan trắc hiện nay, nhằm phục vụ các nghiên cứu về dự báo khí hậu và hỗ trợ ra quyết định trong lĩnh vực môi trường.